Unsupervised Multimodal VAD using Sequential Hierarchy

Rameez Ahmad  
Information Systems, Security and Forensics (ISSF) Lab  
University of Michigan - Dearborn  
Dearborn, Michigan 48124  
Email: rameeza@umd.umich.edu

Syed Paymaan Raza  
Information Systems, Security and Forensics (ISSF) Lab  
University of Michigan - Dearborn  
Dearborn, Michigan 48124  
Email: spraza@umd.umich.edu

Hafiz Malik  
Information Systems, Security and Forensics (ISSF) Lab  
University of Michigan - Dearborn  
Dearborn, Michigan 48124  
Email: hafiz@umd.umich.edu

Abstract—In speech processing systems, the performance of the Voice Activity Detector (VAD) is a bottleneck to the whole system. Traditional VADs are solely based on acoustic features. Additional modality in form of visual information is used to make robust VADs. In this paper, we propose a multimodal VAD based on decision fusion between two modalities. Visual VAD (VVAD) decision vectors are interpolated so that logical operators can be applied to both modalities. In order to avoid this interpolation, we suggest a sequential arrangement of both subsystems to achieve a multimodal VAD. The proposed method considerably reduces false alarm rates when compared with performance of standalone audio VAD (AVAD).

I. INTRODUCTION

In recent years, many speech recognition systems have been proposed which utilize a VAD as a front-end block. VADs are either audio based or video based. Traditional AVADs are based on frame energy, zero crossing rate and entropy. These VADs are sensitive to noisy environment. Though statistical framework for AVADs are more robust relative to heuristic ones, there is still room for improvement. Most of these statistical frameworks also require noise estimates before hand, making these systems semi-supervised. Recently, visual cues have been added to the picture to improve the performance of VAD. VVADs have considerably reduce the load on AVADs by exploiting the combination of audio and video features for making decisions. Another advantage of this combined system is that even when separate systems perform poorly, the combined system can give relatively better performance.

Combination of such sub-systems is achieved using fusion of data. Two fusion algorithms have been proposed in literature. One is early fusion, based on extracting features and then combining them to form one feature representing both audio and video information. The second is late fusion, based on decisions made by both subsystems. Sequential ordering of subsystems can be considered as decision fusion also. Feature fusion is considered to be more efficient overall [1] but the feature extraction process is computationally expensive which requires considerable training of the system. On the other hand, decision fusion performs better on average and is computationally less expensive.

In [2], the authors have proposed decision fusion by placing VVAD followed by a AVAD. However, the method proposed is supervised. Moreover, the feature extraction is based on PCA, which is computationally expensive. In [3], another such method is discussed which is also based on decision fusion. It places unimodal detectors to capture both audio and video data, and then models these using temporal variation of audio and video features using Hidden Markov Models (HMM). However, this system is also supervised since it requires training of audio data. Some other systems in the literature have also proposed multi modal VAD systems. One such system is [4], in which the author achieves good performance in noisy environments. But, this method requires SNR estimates for calculating fusion weights, making it dependent on noise.

In this paper, we propose an unsupervised multi-modal VAD system consisting of Audio and Video VADs. It uses a sequential fusion algorithm to combine the results. Although such systems have been proposed in the past, those systems generally require training. In [5], the author claims AVAD in [6] performs very well. We have used the same AVAD sub-system in our system also, which uses logarithmic energies of a sub-band as features to detect audio presence. For VVAD, we use the method presented in [7], which uses mouth region intensities as visual feature to detect visual speech. Since both sub-systems have different frame rates, decision fusion of these systems is a challenging task. For this reason, we have placed these sub-systems sequentially, VVAD followed by AVAD. Using this scheme, we recorded audio and video in an office environment with ambient lighting conditions and background acoustic noise. With these settings, we were able to achieve considerably good detection rates and low false alarm rates.

II. PROPOSED SYSTEM

A. Visual VAD

As mentioned earlier, we have used the method presented in [7] to detect visual speech. This is because while a person opens his mouth for speaking, there is a dark cavity, which causes a drop in the average intensity of the frame. This average intensity is then modeled using bimodal Gaussian Mixture Model (GMM). In this bimodal GMM, there are two components: speech and silence. Both are modeled using a gaussian with different means and variances. Naturally, the mean of silence is higher than speech while the variance of silence is lower than that of speech.

Once the two components are modeled, a decision boundary is evaluated between them, which is the intersection of
A probabilistic approach is used to classify frames, conditioned on mean, variance and weights of the data. These parameters form a parameter set, which is estimated using the Expectation Maximization (EM) algorithm. In the EM algorithm, constraints are also set. These make sure that the algorithm converges. For example, a minimum mean constraint is set because when a person is not speaking, there should be an artificial component representing visual speech, otherwise, we will be modeling unimodal data with biomodal distribution. This artificial component is set at a distance from the noise component. Additional weight constraint is also included.

For the first M samples of the signal, GMM is initialized using the EM algorithm, including constraints. These M samples are then classified using the decision boundary. This is called the initialization phase. After this, for each incoming sample, the parameter set and decision boundaries are recomputed. And then, the sample is classified as either noise or speech based on the boundary. This is called the updating phase.

It is important to note that, windowing has also been used in the system. Certain window sizes and step sizes are tested on the data set, and then optimal values are calculated. Windowing method ensures that transitory errors are avoided and there is no loss of information during transitions.

### B. Audio VAD

In [5], the authors compared performance of some famous AVADs proposed in literature. The authors divided the algorithms in heuristic and statistical subgroups and briefly described distinguishing features of each algorithm. The experimental results in [5] show that the algorithm proposed by [6] achieved an overall consistent performance among others even in different types of noise. For this reason, we have based AVAD on [6]. The rest of paper will address [6] as Ying’s algorithm. Unlike most AVADs, Ying’s algorithm is totally unsupervised.

Ying’s algorithm aims to distinguish between speech and non speech frames. It uses subband logarithmic energy as an acoustic feature to decide whether there was speech in the frame. Logarithmic energy of a particular Mel subband, out of N subbands, is calculated using logarithmic value of magnitude sum of a particular subband. The logarithmic energy is assumed to follow a bimodal Gaussian distribution. The average energy of a non speech frame will be lower than average energy of a speech frame for a particular subband. This mean difference is exploited to differentiate between speech and non speech. Since noise is stationary as compared to voice, the variance in energy of non-speech will be less than that of speech. The parameter such as mean variance and weight of distribution is estimated using the EM algorithm. By using these parameters, a decision boundary is calculated for each frame so that if the frame energy is below the decision boundary it is labeled as non speech for a subband and vice versa.

There are two phases of the EM algorithm, one is initializing the parameters using M frames and the other is updating the parameter set using incoming new data. The authors proposed a novel approach for the updating phase by assuming that parameter set is varying slowly. The decision boundary tracks the variation of logarithmic energy of each subband. For low SNR situations, Ying’s algorithm adds some constraints about the mean difference between a two component Gaussian. It makes an artificial speech component of a two component GMM, which is at a fixed distance from non-speech mean. This mean constraint leads the EM algorithm to diverge which is dealt by adding a weight constraint. When a weight constraint is activated, it does not let the parameters to be updated. Voting between the N subband’s decision is done to get a final decision about a frame.

Though Ying’s algorithm gives remarkably good results as compared to other AVADs, there are some inherent limitations. The algorithm performance decreases if a new noise source is introduced in the environment. The decision boundary fails to track the energy of subband, which causes the system to give false alarms, which can be compensated by using VVADs. Fig. 1 shows a typical audio sequence and Fig. 2 shows the corresponding tracking and decision boundaries. The effect will be discussed in following section.

### C. Data Fusion

Once we have the decision vectors from both subsystems (AVAD and VVAD), the visual speech frames from the VVAD are sent over to the AVAD. Before doing this, since both
systems have different frame rates, a mapping function is needed to map the video frames to corresponding audio frames. This is the reason a sequential arrangement of subsystems was used. The mapping depends on the window size and step size used in each sub-system. First, the video sequence is passed to the VV AD, which outputs a decision vector. From this decision vector, all the transitions are marked and stored. Note that this decision vector is windowed, which means that it's length is less than that of the original video vector. To solve this, these transitions are mapped to the original sequence by using window and step sizes. Once that is done, the transitions are mapped to the original time stamp by dividing the sequence with frames per second (FPS) of the video.

Knowing transitions (in seconds) that our VV AD sub-system gave as an output, the next task is to map these time transitions to the corresponding frame in the audio sequence. This is done by multiplying the time transition by the sampling rate of the audio. The selected audio frames (between transitions where our VV AD indicate the person spoke), is given to the AVAD sub-system as an input. The AVAD system then gives a decision, and these decisions are stored.

This process is then repeated for all video transition frames which the VV AD sub-system indicates as visual speech. All such transitions are mapped and given as an input to the AVAD system which makes decisions. As mentioned earlier, all these decisions are stored. At the end, the system has a final decision vector accumulated by the AVAD sub-system, but based on the VVAD sub-system. We will use this decision vector for our performance evaluation in the next section. It is important to note that if the VVAD sub-system indicates certain frames as silence, those frames are disregarded and not passed to the AVAD sub-system. This is because we base all our decisions on the VVAD since it is more robust to acoustic noise. Fig. 3 shows our system block diagram.

III. PERFORMANCE EVALUATION

A digital camera with in-built microphone was used to record video as well as audio. The recording was done in an office setting with ambient lighting conditions and background acoustic noise. Following, are the exact dataset specifications:

Table 1: Dataset specifications

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Sampling Rate</td>
<td>8 Khz</td>
</tr>
<tr>
<td>Original Sampling Rate</td>
<td>44.1 Khz</td>
</tr>
<tr>
<td>Audio Quantization</td>
<td>16 bits</td>
</tr>
<tr>
<td>Video Frame Rate</td>
<td>25 FPS</td>
</tr>
<tr>
<td>Video Format</td>
<td>.mpg</td>
</tr>
<tr>
<td>Video dimension</td>
<td>640 x 480</td>
</tr>
<tr>
<td>Video Type</td>
<td>frontal</td>
</tr>
<tr>
<td>Distance of subject</td>
<td>0.56 m</td>
</tr>
<tr>
<td>Number of individuals</td>
<td>14</td>
</tr>
<tr>
<td>Total Number of frames</td>
<td>19573</td>
</tr>
<tr>
<td>Utterance</td>
<td>random</td>
</tr>
</tbody>
</table>

For performance evaluation measures, we define $P_D$ as the ratio of correctly detected speech frames to the total number of speech frames. $P_F$ is defined as the ratio of silence frames, which were misdetected as speech to the total number of silence frames.

Table 2 shows the parameter values we used for evaluating detection and false alarm rates. Table 3 shows the performance we achieved by fusion as compared to Ying’s algorithm. Note that although Ying’s algorithm gives good detection rates, the false alarm rate is relatively high. On the other hand, our algorithm has a decent detection rate while giving a low false alarm rate. This is also evident by Fig. 4, 5, 6 where it can be seen that fused decisions are closer to the marked data as compared to AVAD decisions. Hence, the false alarm rate is relatively low.

Table 2: Parameter values

<table>
<thead>
<tr>
<th>Parameter</th>
<th>AVAD Parameters</th>
<th>VVAD Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window Size</td>
<td>160</td>
<td>14</td>
</tr>
<tr>
<td>Step Size</td>
<td>80</td>
<td>7</td>
</tr>
<tr>
<td>$M$</td>
<td>60</td>
<td>72</td>
</tr>
<tr>
<td>Subbands</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.0083</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Performance evaluation

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$P_D$ (%)</th>
<th>$P_F$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ying’s algorithm</td>
<td>85.32%</td>
<td>15.56%</td>
</tr>
<tr>
<td>Fused</td>
<td>80.17%</td>
<td>4.06%</td>
</tr>
</tbody>
</table>

Fig. 3: System block diagram

IV. CONCLUSION

In this paper, we have proposed an unsupervised multimodal VAD. The performance of the proposed system is
dictated by a VVAD as it is robust to acoustic noise. Decision fusion was selected as the fusion strategy as it is difficult and computationally expensive to do feature fusion. This two stage approach has allowed us to avoid interpolation of one modality decisions. Additionally, the two stage system has also caused significant decrease in false alarm rates when compared with the AVAD. Moreover, the fused system is also far less computationally expensive than the AVAD as, AVAD is activated only if the VVAD detects visual speech.

V. REFERENCES


